DOCUMENT RESUME

ED 273 669 TM 860 519

AUTHOR Lin, Miao-Hsiang
TITLE The Impact of Time Limits on Test Behaviors.
PUB DATE Apr 86
NOTE 42p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Age Differences; Computer Assisted Testing; *Construct Validity; Difficulty Level; Higher Education; Mathematics Tests; *Predictive Validity; Research Design; Sex Differences; *Testing; Test Items; *Timed Tests; Time Management; Verbal Tests
IDENTIFIERS Graduate Record Examinations

ABSTRACT

Specific questions addressed in this study include how time limits affect a test's construct and predictive validities, how time limits affect an examinee's time allocation and test performance, and whether the assumption about how examinees answer items is valid. Interactions involving an examinee's sex and age are studied. Two parallel forms of both sentence completion and non-graphic quantitative items were administered by a computer monitor to 80 volunteer graduate students. They were randomly assigned into two groups in which the form of the test and the order of the timed and untimed conditions were counterbalanced. Except for one form of the quantitative tests, little evidence was found to suggest that a test's construct and predictive validities were jeopardized by time limits. Items whose difficulty values were affected the most by test speededness appeared to be measuring higher- rather than lower-order cognitive abilities. It was found that most students answered items sequentially and did not omit items once they were considered. The caveats to the above findings are discussed. Several additional findings as well as suggestions for further study are provided. (Author/JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED273669

The Impact of Time Limits on Test Behaviors

Miao-Hsiang Lin, Ph.D.

Cornell University
Education Department

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. H. Lin

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 860 519

Abstract

Specific questions addressed in this study include how time limits affect a test's construct and predictive validities, how time limits affect an examinee's time allocation and test performance, and whether the assumption about how examinees answer items is valid. Interactions involving an examinee's sex and age are studied. Two parallel forms of both sentence completion and non-graphic quantitative items were administered by a computer monitor to 80 volunteer graduate students. They were randomly assigned into two groups in which the form of the test and the order of the timed and untimed conditions were counter balanced. Except for one form of the quantitative tests, little evidence was found to suggest that a test's construct and predictive validities were jeopardized by time limits. Items whose difficulty values were affected the most by test speededness appeared to be measuring higher- rather than lower-order cognitive abilities. It was found that most students answered items sequentially and did not omit items once they were considered. The caveats to the above findings are discussed. Several additional findings as well as suggestions for further study are provided.

The Impact of Time Limits on Test Behaviors

The purpose of this study is to provide a better understanding of the effect of time limits on test behavior. A score on a timed test has been recognized as a function of level (ability) and of speed (rate of work) components. When scores on a time-limit test designated to measure level component are contaminated by speed component, the test's construct validity is jeopardized. Baxter (1941), as well as Davidson and Carroll (1945) reported that the factor contents of a timed score were different from those obtained under an untimed condition. Lord (1956) and Mayer (1952) indicated that the score variance of a timed test contained separate portions attributable to speed and level ability. Gulliksen's (1950) and Cronbach's (1951) mathematical models also showed how speed affects a test score's variance and inflates split-half and Kuder-Richardson reliabilities. Studies suggesting that speed is independent of level and that speed rather than level is more related to examinees' temperament have additionally shown the negative effect of time limits on test performance (Wesman, 1960; Mollenkopf, 1960; Boag and Neild, 1962; Terranova, 1972).

Several studies, however, have presented the positive aspects of time limits. It has been found that time limits did not always introduce speed variance even when they were severely stringent (Lord, 1956). In addition, a speed factor can represent ability rather than response set. Bennett's (1956) and Davis's studies (1957) showed that speed performance was intimately related to the students' knowledge when measured by

vocabulary and comprehension tests. Several studies, moreover, indicated that shorter time limits could yield validity coefficients either as high as or higher validity coefficients than longer or no time limits (Baxter, 1941; Mayer, 1952; Allision, 1953; Lord, 1956; Kendall, 1964).

The presence of these contrary results suggests not only that a better understanding of the role of speed in testing is needed, but also that time limits should not be prescribed arbitrarily. The study, therefore, is motivated by a desire to investigate how speed operates and how it relates to test function in a simulated measurement situation that accounts for the interaction of individual differences with varying complexities of test items.

FRAMEWORK

Construct Validity. Previous studies on the impact of time limits on tests' construct validity regarded the untimed score as the criterion of level ability. They all converged upon comparing a timed score with an untimed score. Those studies, however, proposed different rationales and approaches. For example, Baxter's correlation coefficient (1941), Davidson and Carroll's regression technique (1945), Cronbach's Tau index (1951), and Lord's determinant analysis (1971) are all proposed to measure test speededness. To fulfill one of the study's purposes that examines the nature of speed during the test-taking process, this study adopts Davidson and Carroll's regression rationale instead of Cronbach's and Lord's procedures, which are

only designed to measure the influence of speed. Davidson and Carroll's rationale is chosen instead of Baxter's, which is apt to obtain a spurious correlation because he erroneously treated timed scores as speed scores. However, a modified procedure, designed to eliminate problems involved in Davidson and Carroll's approach, is developed. To avoid obtaining a potential spurious correlation and practice effect, the timed and untimed scores are obtained from administering parallel tests, rather than the same test, to examinees. Although time taken to complete the untimed test is also defined as the speed score, the speed scores are the total time that an examinee is willing to spend over the entire untimed test, rather than the sum of time spent on each item the first time it is attempted. Moreover, considering that the time allocation on items is more accurately measured and recorded by computer, the test administration is controlled by a computer.

A standard time limit may generate different levels of time constraint. Some examinees feel extremely rushed, others feel a little rushed, and still others don't feel rushed at all. To describe the extent to which speed is more against students who report greater time pressure, the comparison of an untimed score with a timed score is conducted not only for the group as a whole, but also for the groups expressing the same feelings of being rushed. Three questions are to be examined:

1. To what extent are the scores under the timed condition determined by speed and level components, respectively?
2. Does such determination hold true for groups of students who profess the same degree of being rushed?

3. Does such determination hold true for examinees regardless of differences in sex and age?

Predictive Validity. Kendall (1964) stressed the importance of a time limit as a variable for determining a test's predictive validity. He suggested that a time limit was legitimate so long as it can yield a validity equal to, or higher than, an untimed limit. The present study examines any difference in a test's predictive validity when measured under both timed and untimed conditions. Moreover, the sample characteristics of age and sex are also examined in relation to the two validity coefficients. Since testing materials were selected from formal GRE tests, students' previous GRE scores served as the criterion variable. Two questions to be analyzed are:

4. Do GRE scores from a previous, standard administration correlate more highly with untimed scores than with timed scores?
5. Is the magnitude of the two correlations different for students of different age and sex?

Time Allocation on Test Items. Considering that a test is composed of items varying in their difficulty as well as requiring different types of reasoning, the study conducts a content analysis of test items in relation to the effect of time limits. This facilitates understanding the process of the effect of time limits with relation to item difficulty. Therefore, the study obtains, in order to compare time and untimed conditions, the average time spent on an item and the proportion of students

getting the item right. Questions to be answered for these problems are the following:

6. Is there any discrepancy in the average time allocation on items between the timed and untimed conditions?
7. Is there any discrepancy in the difficulty indices of items between the timed and untimed conditions?

Test-Taking Behavior. Gulliksen's estimate of test speededness is based on the assumption that examinees proceed orderly through the test items; the estimate defines the unattempted items as indicators of speededness. Criticisms have been raised against his assumption in spite of its advantage of requiring only one test administration. Since the examinees' test-taking behaviors can be recorded exactly by computer, this study obtains empirical data to test the concerns addressed by the questions below:

8. Are the students' test-taking behaviors consistent with Gulliksen's (classical) assumption?
9. Upon approaching time limits, do some students fill up blanks with random answers?

In summary, the study examines the effect of time limits on tests' construct and predictive validities. Moreover, capitalizing upon information obtained by computer test administration, the study examines the effect of time limits on item performance and tests the validity of Gulliksen's assumption.

METHODOLOGY

Computer Administration. A PASCAL program with user-friendly algorithms was developed for test administration. A computer screen, during the testing period, displayed an answer sheet, time message and direction of item selection along with an item question. All a student had to do was press an item number, hit the return key, and press an answer choice. The test items could be skipped or selected out of order, and the same items were allowed to be called in repeatedly. The duration was measured between when an item was displayed and a response given. If an item was attempted more than once, the final cumulative figure of this duration was the assumed time spent on the item, but the time spent during each attempt was also recorded. When time was up during the timed condition, the screen image was frozen and students were unable to input any key or call up new items from the screen.

Subject and Testing Materials. The 93 participants were volunteer graduate students at Cornell University. They completed a questionnaire giving personal information reflecting the factors (sex, age, and study field) to be examined in the study. The questionnaire also solicited the students' permission to release their Graduate Record Examination (GRE) scores from the graduate school. Thirteen students, whose GRE scores were not accessible, were selected as the pilot group. The rest of the 80 students remained as the Study Group.

The testing materials were items taken from nongrahic quantitative and sentence completion sections of previous GRE

tests released by the Educational Testing Service. A random stratified (by content) procedure was utilized to develop a pair of 20-item parallel forms (A and B) for both the sentence completion and the quantitative test sections to be administered to the pilot group. The testing materials used with the study group consisted of the same tests, each shortened by 4 items.

Purpose and Procedure of the Pilot Study. Objectives of the pilot study were to estimate reasonable time limits to be set in the timed condition and to determine a test length not too tiring for the subjects. The pilot subjects took the 4 tests in the sequence that demanded the two forms of a given type (sentence completion or quantitative) to be completed before the other type was given. To impose time constraints on the pilot subject, they were asked to complete Form B of a given type under rushed directions such as: "You are expected to do every item as quickly and accurately as possible"; "Please do not omit any item and review any item you have done". The corresponding Form A of either type was administered under the untimed condition that allowed the students to proceed through the tests at their own pace.

Results and Decisions of the Pilot Study. A majority of the pilot students reported that answering 80 test items was somewhat exhausting, so test length was reduced from 20 to 16 items of a given form. The mean, standard deviation, and range of time needed to complete the 16 selected sentence completion items (Form B) were 12.3 minutes, 2.7 minutes and 11.2 minutes, respectively. For the quantitative test (Form B) these values

were 23.4 minutes, 6.4 minutes, and 22.8 minutes, respectively. The relatively large values for the ranges and standard deviations imply that the pace for completing the test was different among the students. They also suggested that a standard time limit might impose different degrees of constraints on the students. However, given the priority of the researcher to maximize the generalizability of the research, a single time limit was set rather than separate time limits for each subject. An arbitrary decision was made to set the time in the timed condition for the study group at 75% of the average time taken by pilot subjects under the rushed directions. Consequently, the time limit of 9.2 minutes (75% x 12.3 = 9.2) was set for the sentence completion tests under the timed condition, while 17.6 minutes (75% x 23.4 = 17.6) was set for the quantitative tests under the timed condition.

The fear of different degrees of difficulty involved in the parallel forms generated another decision that both Form A and Form B of a given type were to be administered under timed and untimed conditions (using different subjects, of course).

Research Design of the Study. The design used in the study was developed to balance the effects of test type, test form and order of time condition. Consequently, the 80 graduate students were randomly assigned into 2x2x2 blocks with 10 students in each block. Table 1 illustrates the design. The 10 students assigned to block 1 were tested in the following order:

Form A of sentence completion under the timed conditon was given first;

Form B of sentence completion under the untimed condition was given second;

Form A of quantitative under the timed condition was given third;

Form B of quantitative under the untimed condition was given last.

Insert Table 1 here

The same rule applied to the students assigned in the other 7 blocks, with the two tests of a given type administered before the two tests of the other type. Students assigned to blocks 1, 4, 5 and 8 were aggregated into group 1 since they took Form A of a given type under the timed condition and took the corresponding Form B under the untimed condition. In the same way, students assigned in blocks 2, 3, 6 and 7 were combined into group 2 since they completed Form A under the untimed condition and Form B under the timed condition.

Testing Procedure. The testing materials were administered to 60 students in groups ranging in size from 3 to 8 within a two-hour session in the same computer room reserved for the study. The other 20 students were individually tested at their convenience due to time conflicts with the initially reserved time. The testing directions and procedures were standardized. The students under group administration were informed that the test order with their corresponding time conditions varied for each person. Consequently, it was hoped that students were not

Table 1

The 2x2x2 Block Design Used in the Study

Test Type Administered First	Sentence Completion				Quantitative			
Test Form Administered First	A		B		A		B	
Time Condition	Timed	Untimed	Timed	Untimed	Timed	Untimed	Timed	Untimed
Block	1	2	3	4	5	6	7	8
N	10	10	10	10	10	10	10	10

concerned with the amount of time taken by others to complete the test.

During the practice test, the amount of time permitted to operate the computer keyboard was tailored to the students' needs. Such an arrangement resulted from the investigator's concern that the students' experiences in a computer environment might be different.

Every student received a note listing the time conditions and test order of the four tests. In addition to the note, at the beginning of a new test form the computer monitor displayed directions such as "You can take as much time as you need to complete this test" or "You have to finish this test within 17 minutes". After finishing the last test, each student received a questionnaire asking how rushed they were and how they solved items under timed versus untimed conditions.

RESULTS

Similarity of Both Groups. Among the variables used for determining the equivalence of the two groups were sex, study field, age, and previous GRE verbal and quantitative scores. A chi-square test was used to compare the obtained frequencies of sex and study field in both groups. A t-test on mean differences and an F test on variance were employed for the other variables of age, GRE verbal and quantitative scores. None of the calculated values X^2, t and F were statistically significant. These results suggested that to some degree both groups could be considered equivalent.

Parallelism of Test Forms. To examine the degree to which both test forms were parallel, score means, variances, reliabilities and standard errors of measurement were compared for the test forms when they were administered under the same time condition. Table 2 presents these comparisons. The differences in the four statistics between Form A and B were negligible. These small difference held true for both time conditions and for both test types. Form A and Form B, therefore, might be considered statistically parallel.

Insert Table 2 here

Results of Questions on Construct Validity. Multiple regression equations were employed to provide evidence about Question 1 concerning the importance of speed and level in determining a timed score. Since speed and level differed in measurement units, the standardized regression coefficients, instead of the raw regression coefficients, are presented with the results summarized in Table 3.

Insert Table 3 here

The multiple correlations for which test scores are predicted from level and speed factors are statistically significant in all cases. The beta value for the level factor is significant for both types of tests, whereas the beta value for the speed factor is significant only in the prediction of scores

Table 2

Means, Standard Deviations, Alphas and Standard Error of Measurement of Forms A and B for Both Test Types

		Timed Condition		Untimed Condition	
		Form A	Form B	Form A	Form B
Test Type	Statistics	Group 1 (n=40)	Group 2 (n=40)	Group 2 (n=40)	Group 1 (n=40)
	Mean	11.65	11.87	12.28	12.28
Sentence	S.D.	2.33	2.39	2.09	2.22
Completion	Alpha	0.59	0.63	0.52	0.58
	Se	1.49	1.45	1.44	1.43
	Mean	12.08	11.65	13.75	13.15
Quantitative	S.D.	2.72	2.39	2.09	2.41
	Alpha	0.71	0.67	0.66	0.68
	Se	1.48	1.38	1.21	1.35

Table 3

Sentence Completion and Quantitative Test Scores Obtained Under Timed

Conditions Regressed on Factors of Level and Speed

Terms in the Equation	Y (timed score)	+ B1X1 (untimed score as level)	+ B2X2 (completion time as speed)	+ A
Test Type	Sentence Completion		Quantitative	
Group	1	2	1	2
Dependent Variable	Form A (timed)	Form B (timed)	Form A (timed)	Form B (timed)
Multiple R	0.547	0.689	0.493	0.733
R Square	0.298*	0.474*	0.243*	0.538*
Regressor				
X1 Beta	0.549*	0.629*	0.465*	0.571*
X2 Beta	0.035	-0.113	-0.341*	-0.446*

Note: X1 is for untimed scores (level).
X2 is for completion time under untiomed condition (speed).

*p<0.05

on the quantitative test items. Although the beta value for the level factor is higher than that for the speed factor in the quantitative test, the absolute value ratio of the former to the latter is only either 1.36 or 1.28.

The intent of Questions 2 and 3 is to explore whether the effect of level and speed on timed scores varied with examinees' characteristics such as feeling of being rushed, sex and age. Table 4 displays the frequencies corresponding to these three variables.

Insert Table 4 here

Statistically, the core interest of these two questions is in comparing the equality of the slopes of the regression lines computed separately for each category of the variables of self reported feelings of being rushed, sex and age. To determine whether the use of common regression coefficients was appropriate for those separate regression lines of a variable, an overall regression equation included timed scores as the dependent variables and set the following groups of regressors to enter the equation sequentially in the order as they are listed below:

1. Untimed scores as level (X1), completion time under untimed condition as (X2)
2. Effect coding vector(s) of the categorized variables (V1 only or V1 and V2)
3. Product terms of level and speed with coded vectors.

Table 4

Frequencies of Feeling of Being Rushed, Sex and Age for Each Group and Each Test Type. Data are for the Timed Administration Only.

Variable	Group	Test Type					
		Sentence Completion			Quantitative		
		Extremely Rushed	A Little Rushed	Not at All	Extremely Rushed	A Little Rushed	Not at All
Feeling of Being Rushed	1	6	24	10	7	17	16
	2	9	20	11	15	20	5
		Female	Male		Female	Male	
Sex	1	18	22		18	22	
	2	23	17		23	17	
		Age $\leq$ 30	Age $>$ 30		Age $\leq$ 30	Age $>$ 30	
Age	1	36	4		36	4	
	2	33	7		33	7	

Note: Computer coding for the three variables is the following

Feeling of Being Rushed	Sex	Age
1: extremely rushed	0: female	1: equal to or less than 30
2: a little rushed	1: male	2: over 30
3: not at all		

Results corresponding to questions 2 and 3 are listed in Tables 5 and 6. As shown from part 1's of Tables 5 and 6 the R square change of the product terms (X1V1-X2V2) in both groups are minute and not significant at the 0.05 level. This led to the conclusion that use of a common regression line of the three variables is appropriate. The one exception is for the age factor, in which quantitative scores are predicted for students in Group 1 (see Table 6). Therefore, a regression equation with regressors of level and speed only was computed for the two age categories in Group 1 respectively. Results are displayed in part 2 of Table 6.

Insert Table 5 and 6 here

With respect to the sentence completion tests, the betas for level are significant at the 0.05 level while the betas for speed are not significant for both groups (see part 2 of Table 5). That the variance of timed scores of sentence completion tests contained significant level variance but negligible speed variance held true for students regardless of which characteristic (degree of being rushed, sex, or age) was included in the regression equation.

Results for the scores on the timed quantitative tests were less clear cut. Although the beta weight for level was typically higher than that for speed, there was one exception (students in Group 1 less than 30 years old) (see Table 6). The beta weight for level and speed were both typically significant. The two

Table 5

Sentence Completion Scores Obtained Under Timed Conditions Regressed on Level, Speed, and Effect Coding on Feeling of Being Rushed, Sex, and Age

Coding Factor	Feeling of Being Rushed		Sex		Age	
Group	1	2	1	2	1	2
Dependent Variable	Form A (timed)	Form B (timed)	Form A (timed)	Form B (timed)	Form A (timed)	Form B (timed)
Part 1	R Square Change	R Square Change	R Square Change	R Square Change	R Square Change	R Square Change
X1,X2	0.298*	0.474*	0.298*	0.474*	0.299*	0.474*
V1,V2	0.045	0.077	0.000	0.006	0.034	0.000
X1V1-X2V2	0.029	0.013	0.042	0.004	0.004	0.005
Part 2 Regressors in Equation						
	Beta	Beta	Beta	Beta	Beta	Beta
X1	0.567*	0.526*	0.549*	0.624*	0.558*	0.632*
X2	0.081	-0.140	0.034	-0.121	0.045	-0.112
V1	0.096	-0.273	-0.005	-0.077	-0.186	-0.017
V2	-0.254	0.298				

Note: X1 is for untimed scores (level), X2 is for completion time under untimed condition (speed). V1 and V2 are the effect coding vectors.
* $p < 0.05$

Table 6

Quantitative Scores Obtained Under Timed Conditions Regressed on Level, Speed, and Effect Coding on Feeling of Being Rushed, Sex, and Age

Coding Factor	Feeling of Being Rushed		Sex		Age	
Group	1	2	1	2	1	2
Dependent Variable	Form A (timed)	Form B (timed)	Form A (timed)	Form B (timed)	Form A (timed)	Form B (timed)
Part 1	R Square Change	R Square Change	R Square Change	R Square Change	R Square Change	R Square Change
X1,X2	0.243*	0.538*	0.243*	0.538*	0.243*	0.538*
V1,V2	0.256*	0.100*	0.008	0.049	0.091*	0.036
X1V1-X2V2	0.088	0.029	0.012	0.038	0.137*	0.006
Part 2 Regressors in Equation	Beta	Beta	Beta	Beta	Age≤30	Age>30
X1	0.206	0.516*	0.487*	0.560*	Multiple R	0.526* 0.901
X2	-0.010	-0.313*	-0.344*	-0.398*		
V1	-0.658	-0.275	0.094	-0.228	Beta	Beta
V2	0.128	-0.194			X1	0.306 0.806
					X2	-0.507* 0.193

Beta	
X1	0.532*
X2	-0.361*
V1	0.211

Note: X1 is for untimed scores (level), X2 is for completion time under untimed condition (speed), V1 and V2 are the effect coding vectors.
*$p < 0.05$

exceptions were for students in Group 1 for the rushed variable and for the over 30 category in Group 1. In these two cases, neither beta weight reached statistical significance. In order to examine the main effect of age in relation to speed and level, a multiple regression equation without any interaction terms was computed for quantitative tests in both groups. Results show that when age is held constant, the beta weights for speed and level are both statistically significant at the 5% level. For both groups the beta weight of level is higher than that for the speed factor.

Results of Questions on Predictive Validity. The Hotelling-Williams test, conducted for questions 4 and 5, was used to compare the two correlations between test validity with timed scores and with untimed scores. The GRE-verbal score was used as the criterion for the sentence completion test; the GRE-quantitative score was used as the criterion for the quantitative test.

As shown in Table 7, the external validities of the timed and untimed scores are not statistically different under a two-tailed test. This lack of difference is true for both groups and for both test types regardless of sex or age.

Insert Table 7 here

Results of Questions on Time Allocation on Test Items. The intent of questions 6 and 7 is to investigate if extension of

Table 7

A Comparison of the Predictive Validities Under

Timed and Untimed Conditions

			GRE Verbal Aptitude Form A (timed)	GRE Verbal Aptitude Form B (untimed)	GRE Quantitative Aptitude Form A (timed)	GRE Quantitative Aptitude Form B (untimed)
Group	1	N = 40	0.64	0.62	0.63	0.65
	2	N = 40	0.79	0.66	0.64	0.68
Group	1	Female (N = 18)	0.51	0.35	0.68	0.38
		Male (N = 22)	0.73	0.75	0.61	0.84
Group	2	Female (N = 23)	0.78	0.69	0.56	0.52
		Male (N = 17)	0.82	0.65	0.77	0.75
Group	1	Age $\leq$ 30 (N = 36)	0.63	0.64	0.59	0.61
		Age $>$ 30 (N = 4)	0.65	0.64	0.94	0.98
Group	2	Age $\leq$ 30 (N = 33)	0.79	0.65	0.66	0.62
		Age $>$ 30 (N = 7)	0.80	0.81	0.29	0.18

time limits increases time allocation on every item or on some particular items, as well as if the increased time allocation yields higher "difficulty" indices. Content analysis on test items is one core interest shared by the two questions. The data for these research questions are based on the results of two groups, one answering the items under the timed condition and the other (statistically equivalent) group answering the identical items under the untimed condition. A t test and a X^2 test were employed for questions 6 and 7, respectively. In addition, means of time allocation on items and the difficulty index of each item are all displayed in Table 8 for both test types.

Insert Table 8 here

In comparing the average time allocation on items, data of Table 8 confirm the known fact that students spend more time on items under the untimed condition than under the timed condition. However, a trend strongly shows that the untimed condition induces students to spend more time on harder items than on easier ones. With respect to item difficulties (P), and as shown by X^2 values, none of P values of the sentence completion items and only 6 of the P values of the 32 quantitative items are statistically different between the two conditions.

Results of Questions on Test-Taking Behavior. Based on the computer records, it was found that 74 of 80 subjects answered items consecutively in the sentence completion tests, while 75 of

Table 8

Means of Item Allocation on Items and Difficulty Indices of Items in Sentence Completion and Quantitative Tests Under Untimed and Timed Conditions as well as the Corresponding Values for Mean Comparison and Chi-Square Values for P Comparison

Test Type		Sentence Completion								Quantitative							
Test Form		Form A								Form B							
Item	Time Cond.	Mean	t	P	χ^2	Mean	t	P	χ^2	Mean	t	P	χ^2	Mean	t	P	χ^2
1	U	27.98		1.000		24.30		0.950		73.90		0.975		25.73		1.000	
	T	22.30	2.00	0.950	0.51	20.60	1.89	1.000	0.53	56.60	1.66	0.875	1.62	21.43	1.53	0.975	0.00
2	U	30.40		0.925		28.25		0.975		21.47		0.925		23.48		1.000	
	T	21.83	3.01*	1.000	1.38	26.90	0.56	1.000	0.00	19.13	0.79	0.925	0.00	21.02	0.86	1.000	0.00
3	U	15.83		1.000		23.25		1.000		23.95		0.975		30.20		1.000	
	T	12.93	1.94	0.975	0.00	19.60	1.33	0.975	0.00	23.73	0.10	0.950	0.00	25.68	1.59	0.975	0.00
4	U	21.15		1.000		30.80		0.975		53.03		0.925		47.13		0.900	
	T	15.65	2.96*	0.925	1.38	27.08	1.19	0.975	0.00	40.15	1.97	0.925	0.00	47.48	-0.07	0.950	0.18
5	U	32.48		0.950		18.38		0.905		20.28		0.925		49.53		0.975	
	T	29.65	1.98	0.950	0.00	15.60	1.77	1.000	0.00	17.18	1.52	0.875	0.14	49.48	0.01	0.925	0.26
6	U	37.23		0.750		49.50		0.775		37.65		0.950		75.23		0.875	
	T	29.65	2.80*	0.850	0.70	37.98	2.20*	0.800	0.00	39.08	0.23	0.975	0.00	57.73	2.31*	0.850	0.00
7	U	67.18		0.725		25.10		0.900		67.55		0.725		89.78		0.875	
	T	49.28	3.04*	0.625	0.51	26.88	-0.46	0.800	0.88	62.83	0.42	0.575	1.37	69.25	2.13*	0.850	0.00
8	U	63.20		0.900		36.13		0.850		32.50		0.925		56.20		0.925	
	T	49.55	2.57*	0.775	1.46	31.78	1.45	0.775	0.33	25.83	1.57	0.875	0.14	55.73	0.04	0.800	1.68
9	U	30.48		0.925		50.75		0.850		123.77		0.850		147.00		0.900	
	T	23.55	2.01*	0.825	1.03	37.93	2.46*	0.750	0.70	98.33	2.28*	0.750	0.70	130.55	1.44	0.775	1.46
10	U	30.98		0.675		34.33		0.525		48.15		0.850		84.03		0.950	
	T	23.85	2.36*	0.675	0.00	21.03	3.30*	0.600	0.20	44.73	0.79	0.875	0.00	65.15	1.90	0.950	0.00
11	U	36.03		0.700		71.01		0.700		72.20		0.725		76.38		0.800	
	T	28.60	1.97	0.750	0.06	42.88	2.98*	0.725	0.00	63.78	1.29	0.675	0.06	71.15	0.47	0.750	0.07
12	U	71.25		0.625		52.85		0.625		84.13		0.825		315.90		0.775	
	T	44.33	3.91*	0.600	0.00	36.73	3.26*	0.525	0.46	70.67	1.31	0.700	0.10	141.05	5.43*	0.275	18.09*
13	U	58.98		0.750		59.05		0.650		107.70		0.750		145.23		0.800	
	T	39.75	2.43*	0.725	0.00	38.43	5.07*	0.575	0.21	109.15	0.10	0.725	0.00	74.70	5.66*	0.475	7.78*
14	U	38.73		0.475		70.30		0.725		126.37		0.625		118.93		0.825	
	T	26.00	3.67*	0.450	0.00	35.85	6.12*	0.625	0.51	84.50	3.64*	0.475	1.26	43.17	5.44*	0.500	8.05*
15	U	73.25		0.575		39.05		0.575		66.95		0.750		117.45		0.700	
	T	31.85	7.95*	0.350	3.21	25.45	3.55*	0.450	0.80	53.35	1.79	0.500	4.32*	30.00	5.91*	0.375	7.24*
16	U	43.28		0.300		108.63		0.325		112.13		0.450		107.85		0.425	
	T	31.80	2.75*	0.250	0.06	43.73	5.77*	0.300	0.25	86.02	1.79	0.400	0.05	42.00	5.84*	0.225	3.57*

the 80 answered consecutively in the quantitative tests. However, 65% of the 80 subjects reported that they would have been more inclined not to answer items consecutively if the tests had been administered in the traditional paper-pencil format.

To test whether students, as the item limit approached, answered questions randomly without reading and thinking about them, the time spent on the last six questions was examined. Evidence of unconsidered answers would be if a student spent less than 10 seconds on any of the six items. Based on the computer records, in no case was such evidence found. It was found, however, that 15 and 22 students left one or more items blank on the sentence completion and quantitative tests respectively.

DISCUSSION

Sentence Completion Tests. The variance in the timed scores of the sentence completion test seems to be contributed predominantly by the level factor regardless of the examinees' sex or age. The same results even hold true for those examinees who professed to be under extreme time constraints. Little evidence shows that the test's construct validity is jeopardized by a sufficient speed factor (rate of work). This suggests that the sentence completion test (in the timed condition) may be unspeeded. Such a claim can be further supported by the other two investigations that measure the degree of changing mean scores or standings. It has been found that a non-significant mean difference of test scores, as well as a significant correlation, exist between timed and untimed conditions.

The time limits of 9.2 minutes generally accelerates students' normal rate of work. Such accelerating, however, is found to interact with the difficulty level of items. The larger discrepancy of the time allocation between timed and untimed conditions accompanies the more difficult items. Although the untimed condition permits students to answer questions at their own pace, none of the difficulty indices of the 16 items in both forms are significantly increased under the untimed condition. The implication that the "speeded" rate of work does not significantly change the difficulty index might help explain why the timed score variance contains a negligible speed factor.

The time limits of 9.2 minutes did result in 15 of the 80 examinees (19%) not completing all test items. The findings, however, appear to suggest that time limits did not generate a harmful effect against this group of examinees, who were also those who felt more rushed. As discussed before, the timed score variance, due to the speed factor, is negligible for this group of examinees. Moreover, Lin (1985) shows that the verbal GRE aptitude scores, rather than sex or age, correlate with feeling of being rushed. The relationship is such that students feeling more time constraints are found to have lower aptitude scores, which also significantly correlates with both timed and untimed scores. In addition, a scrutiny of the standings on the timed and untimed score distributions for those 15 examinees who did not finish reveals that 12 of these examinees remain at lowest 12 ranks on both tests. These findings seem to confirm Cronbach's warning: "... it is neither substantive nor sufficient to

classify a test as speeded or biased simply if less than all examinees complete all items."

This study shows that a time limit of 9.2 minutes can yield a predictive validity as high as untimed limits. Such evidence is further found for subsamples corresponding to sex and age status. This study, therefore, concludes that when previous GRE verbal scores are taken as a criterion, the imposition of 9.2 minutes on sentence completion tests with 16 items does not decrease the test's validity that can be obtained under no time limits.

Gulliksen's assumption that students will answer items sequentially is only slightly violated. However, this may be more attributable to test administration by computer than to the students' willingness. The inference is based on 65% of the students reporting they would have been more willing to answer items consecutively if tests were administered in the traditional paper-pencil format. This study, therefore, cannot reach any conclusion about the extent to which Gulliksen's assumption is violated.

Because of computer administration, this study was able to distinguish items deliberately left blank (considered and omitted) from items omitted for lack of time. In no case did a student skip over an item without first considering it, thus supporting Gulliksen's assumption used in computing his speed index.

This study finds that more than 80% of the examinees (the combined group) reached 100% of the test items. By the method of

ETS's operational practice (Swineford, 1974; Donlon, 1973) the sentence completion test is unspeeded. In this case, the ETS method presents results similar to those of the multiple regression method.

One implication related to computer test administration is its potential to maximize the validity of Gulliksen's index and ETS's practice in measuring test speededness. When computer algorithms are designed to only allow examinees to review items having been answered and to take items in order, Gulliksen's assumption won't be violated easily.

Quantitative Tests. The time score variance of the quantitative tests is found to contain a significant portion attributable to the speed factor. The effect of the level factor is larger than that of the speed factor, but the latter is not negligible. The imposition of a 17.6 minute time limit on a quantitative test with 16 items appears to produce a test that is at least partially speeded. Such a claim is strongly supported by the group who took Form B under the timed condition and Form A under the untimed condition. However, for the other group who took the reverse forms and time conditions, the claim is less strongly supported, because of two exceptions (see Table 6).

The time limits of 17.6 minutes imposed on Form A appears to approximate the students' normal rate of working since none of the differences between average time allocation on items are statistically significant, except for items 9 and 14. The effect of time limits also appears not to jeopardize test scores since none of the differences in item difficulty values are

significant, except for item 15. A non-significant difference between mean test scores of Form A under timed and untimed conditions supports this suggestion. Therefore, Form A under the time limit seems to be unspeeded.

When similar comparisons are employed for Form B, the item limit of 17.6 minutes appears to interfere with the students' normal rate of work and their test scores, since more items display a statistically significant difference. A significant mean difference of total test scores on Form B under timed and untimed conditions ($t = 4.19$, $p < .005$) also strengthens this suggestion. Thus, Form B under the time limit appears to be speeded.

By ETS's operational practice for detecting test speededness, Form A is unspeeded (since 88% of examinees finished all test items) while Form B is speeded (since 58% of the examinees finished all test items).

With respect to the effect of time limits on the predictive validity of tests, the time limits of 17.6 minutes can yield a validity as high as no time limits. Such equity does not change by age and sex status. This result suggests that the time limits do not jeopardize a test's predictive validity.

Caveats. The time limits used in this study had little effect on the tests' construct and predictive validities. Results of this study did not show the interaction between the effect of time limits in relation to sex and age. These results, however, might be more attributable to the specific sample and measurement situation of this study. The following caveats might

otherwise decrease the validity and generalizability of such results.

The untimed condition might be, in fact, only a relative untimed condition because examinees might impose time limits on themselves when they were tested under the untimed conditions. The cases where examinees lack motivation to do well on tests or compete to finish tests earlier than other students could reduce the effect of extension of time limits on increasing test scores.

Although the research design of this study was developed to balance the order effect of time condition, the degree to which such balance was achieved is unknown. One could suspect that if examinees did impose time limits on themselves the interaction effect between test performance and test order could confound the results.

The test administration by a computer monitor that did not allow examinees to see the whole picture of test content might jeopardize an ideal time allocation on test items under timed conditions. In such case examinees might have gotten stuck on a particular item and left more items blank. Results of this study might also be positive or negatively confounded by the degree of parallelism obtained for both test forms. The homogeneous group of volunteer students at Cornell and the subsequent homogeneous and high difficulty indices limit the generalizability of such results to the other sample.

CONCLUSION

The time limit used for the sentence completion test might not be stringent enough for the purpose of studying effect of time limits. However, provided the effects of the above caveats are minimal, setting 9.2 minutes on a 16-item sentence completion test could be justified, as the time limit does not jeopardize tests' construct and predictive validity. Moreover, since the structure in sentence completion tests is unique among items, one could infer that when the time limit allows an average of 30-35 seconds per item, test speededness could be negligible, at least for subjects with similar background to those in this study. The study also suggests that it is unwise to impose the same time limits on parallel forms of a test just because two forms have the same test length and measurement statistics (i.e. average difficulty level, reliability coefficients, and standard error of measurement).

It seems that the effect of time limits interact with type of construct to be measured. Such a claim is based on the fact that items whose difficulty indices are significantly changed measure reasoning ability while most items whose difficulty indices remain unchanged measure lower-order abilities. The effect of time limit might be more significant for higher-level cognitive abilities since they demand more thorough thinking process. One could infer the effect of time limits on sentence completion tests is minimal, because sentence completion tests measure large portion of recall ability.

A time limit that results in some examinees leaving items blank can lower the examinees' scores, or change examinees' standing in the group, or both or neither. Results of the sentence completion tests and of Form A of the quantitative tests are such cases where the time limits do not change scores and standings very much. Although Form B of the quantitative tests, with its time limit of 17.6 minutes, proved to be speeded by the methods employed in this study, whether or not the speededness also alters the examinees' standings is unknown. (Only 7 of 17 examinees with items left blank remain the lowest ranks on both Form A and Form B. However, a significant correlation exists between the two forms.) By Cronbach's definition (1951) or by the purpose of norm-reference tests, if the time limit of 17.6 minutes did not change examinees' standings, Form B is unspeeded. By the purpose of criterion-reference tests, Form B is legitimately claimed as speeded since the time limit did decrease the examinees' scores. Therefore, it seems that one can conclude that a time limit that lowers examinees' scores and standings is definitely speeded. A time limit that lowers either scores or changes relative standings is speeded or unspeeded, depending upon which definition of speededness and which testing purpose one is inclined to emphasize.

Suggestions for Further Research. Effect of time limits interacts with time limit conditions, testing materials and sample characteristics. Any experimental study should take the importance of these three dimensions as variables into consideration. For example, the time limit condition should vary

including several levels with different time constraints; testing materials should be examined in terms of different cognitive processes such as recall vs. reasoning or mechanical vs. intellectual; and the sample should be heterogeneous in sex, age and ability level.

When studying effect of time limits, test scores under an untimed condition have been considered as "standard" for comparing with those obtained under a timed condition. It has been found by this study that near 50% of examinees professed to be more inclined to solve problems backwards under untimed conditions. The desired construct to be measured by the test could be interfered with by solving problems backwards, particularly for quantitative questions. Therefore, studies that examine how untimed conditions jeopardize a test's construct validity and how that interacts with testing materials should be further explored.

When measuring test speededness, a research design should weigh between the use of same test and parallel forms for different item conditions. If the choice is the same test, to eliminate a part-whole relationship and practice effect, time elapse between timed and untimed conditions is needed. If the choice is the parallel form, the tests should be long enough to be similar in item statistics and content validity. Moreover, when a research design with varying time limit conditions has more than two levels, it is important to weigh between a cross-random sample design and a repeated-within sample design. The former can gain students' cooperation and overcome practice and

fatigue effect. However, the equivalent samples are hard to obtain.

References

Allison, P.B. (1953). Battery validity as a function of time devoted to each test. Educational and Psychological Measurement, 12, 288-295.

Baxter, B. (1941). An experimental analysis of speed and level in an intelligence test. Journal of Educational Psychology, 32, 285-296.

Bennett, G.K., & Doppelt, J.E. (1956). Item difficulty and speed response. Educational and Psychological Measurement, 16, 494-496.

Boag, A.K., & Neild, M. (1962). The influence of the time factor on the scores of the Triggs Diagnostic Reading Test as reflected in the performance of secondary school pupils. Journal of Educational Research, 55, 181-183.

Cronbach, L.J., & Warrington, W.G. (1951). Time limit tests: Estimating their reliability and degree of speeding. Psychometrika, 14, 167-188.

Davidson, W.M., & Carroll, J.B. (1945). Speed and level components in time limit scores: A factor analysis. Educational and Psychological Measurement, 5, 411-427.

Davis, F.B., & Davis, C.C. (1957). Davis reading test. New York: The Psychological Corporation.

Donlon, T.F. (1973, Fall). Establishing appropriate time limits for tests. Paper presented at the meeting of the Northeast Educational Research Association, Ellenville, NY.

Gulliksen, H. (1950). Theory of mental test. New York: John Wiley and Sons.

Kendall, L.M. (1964). The effects of varying time limits on test validity. Educational and Psychological Measurement, 24, 789-800.

Lin, M.H. (1985). The impact of time limits on test behaviors. Unpublished doctoral thesis, Cornell University, Ithaca, NY.

Lord, F.H. (1956). A study of speed fators in tests and academic grades. Psychometrika, 21, 31-50.

Lord, F.H. (1971). Testing if two measuring procedures measure the same psychological dimension. (RB-71-36). Princeton, NJ: Educational Testing Service.

Mayer, C.T. (1952). The factorial composition and validity of differentially speeded tests. Psychometrika, 17, 347-352.

Mollenkopf, W.G. (1960). Time limits and the behavior of test takers. Journal of Educational and Psychological Measurement, 20, 223-230.

Morrison, E.J. (1960). On test variance and the dimensions of the measurement situation. Educational and Psychological Measurement, 20, 321-250.

Swineford, F. (1974). Test Analysis Statistical Report SR-7406. Princeton, NJ.

Terranova, G. (1972). The relationship between test scores and test time. The Journal of Experimental Education, 40, 81-83.

Wesman, A.G. (1960). Some effects of speed in test use. Educational and Psychological Measurement, 20, 267-274.

END

U.S. DEPT. OF EDUCATION

OFFICE OF EDUCATIONAL RESEARCH AND IMPROVEMENT (OERI)

ERIC

DATE FILMED

JAN_27_1987